

广东--公共交通大数据竞赛

公交线路客流预测

Andy

2015.12

回顾

特征整理

以全部时间数据为研究对象，在基础特征上对不同性质的特征进行整理，对连续性特征进行离散化。

61.27

使用基本特征

将所给数据进行数据清洗后，统计初步特征，带入LR/RF模型中得到最初的结果

75.03

模型升级

尝试采用ensemble的方法替换原有的单一模型；模型参数的调优；针对整体数据使用模型。

78.12

补充假日特征集

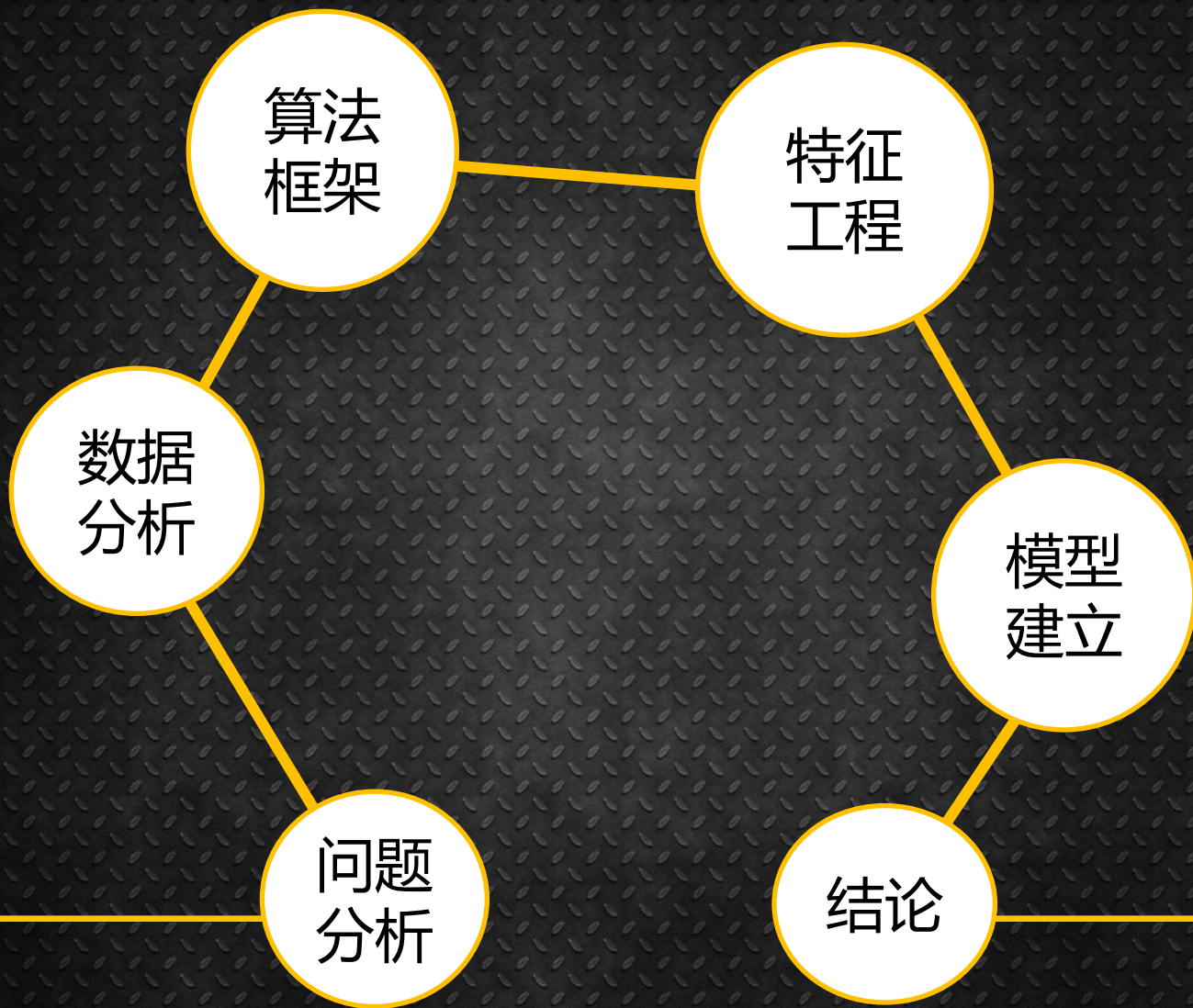
针对节假日的特征集，适当的用weekend特征去拟合（例如周末可以当作节假日）；节假日长度；节假日起始标志等。

80.66

整体调优

调整采用的训练集合，舍弃部分时间数据；根据预测图像调整部分特征，进行整体调优。

79.61





问题分析

问题分析

分类&回归?

任务：使用2014年8-12月的公交刷卡数据和天气信息，预测2015年1.1-1.7的各个时段的乘车人次。



问题分析

数据分析

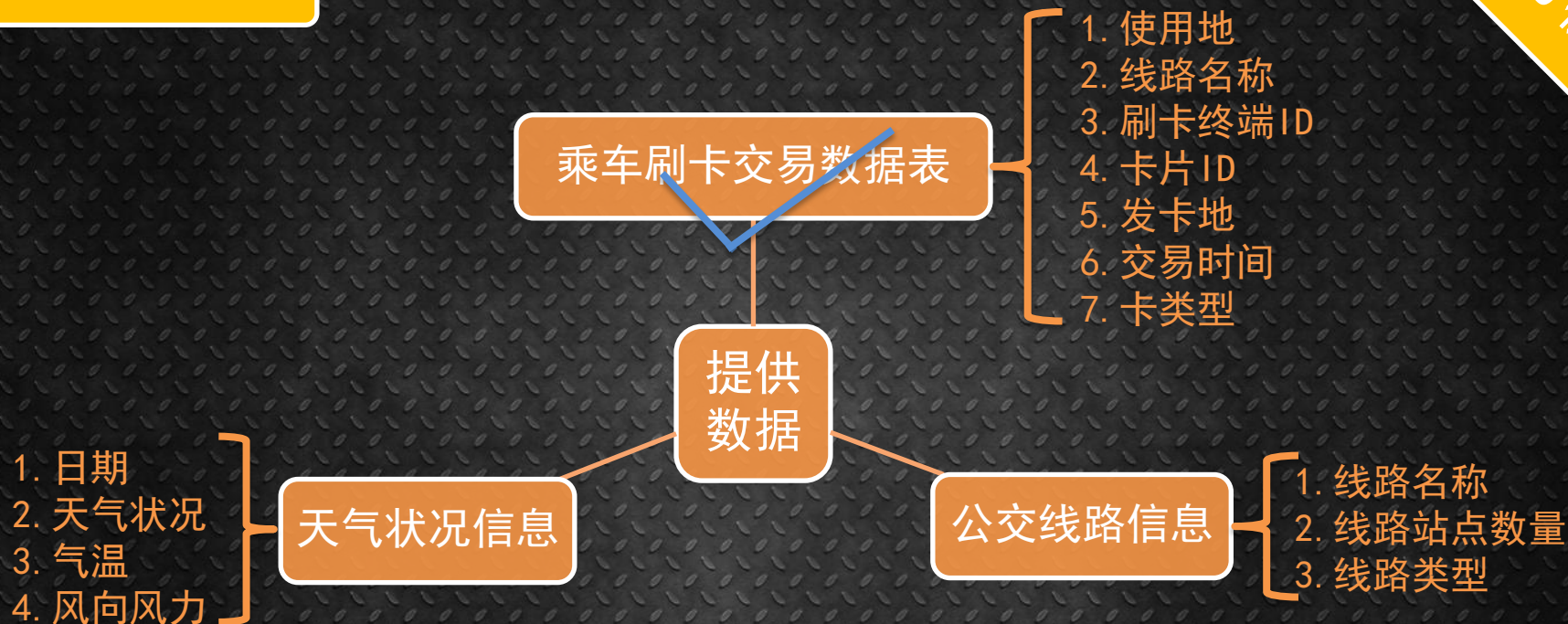
算法框架



数据
分析

数据构成

数据分析



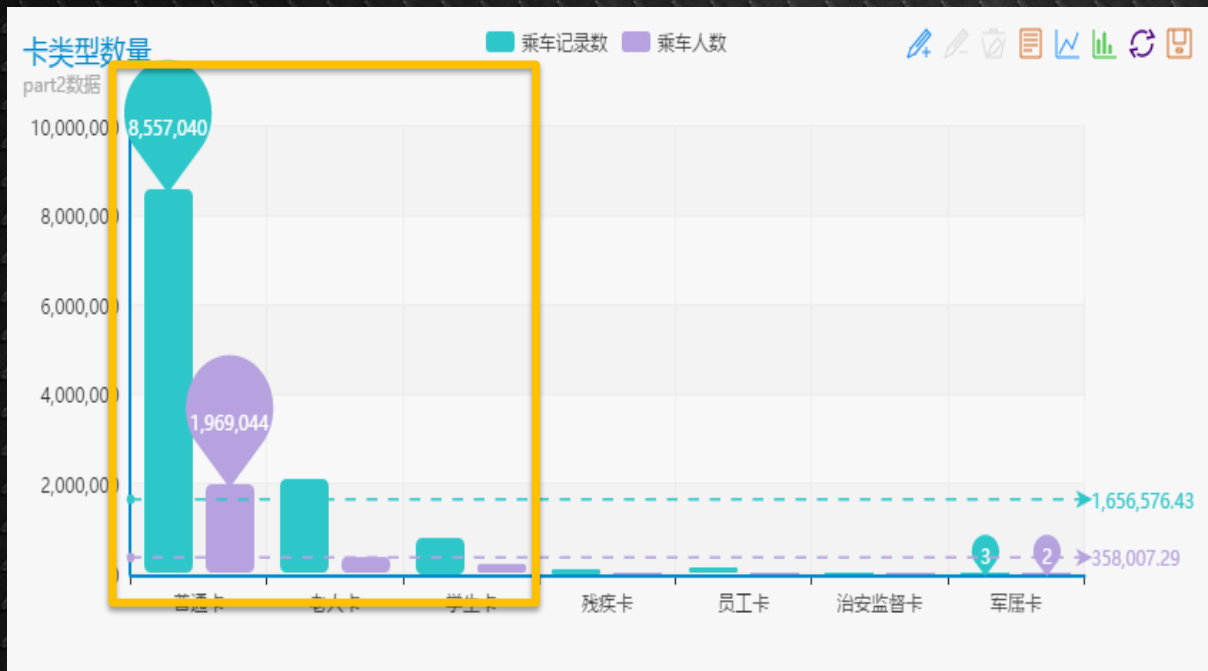
问题分析

数据分析

算法框架

1、乘车类型统计

Part2 --- 乘车总记录数：11596035
乘车总人数：2505883



卡类型数量

普通卡：8557040，1969044
老人卡：2073941，333463
学生卡：786480，175751
残疾卡：79157，13421
员工卡：96925，14002
治安卡：2489，368
军属卡：3，2

分类型乘车记录统计

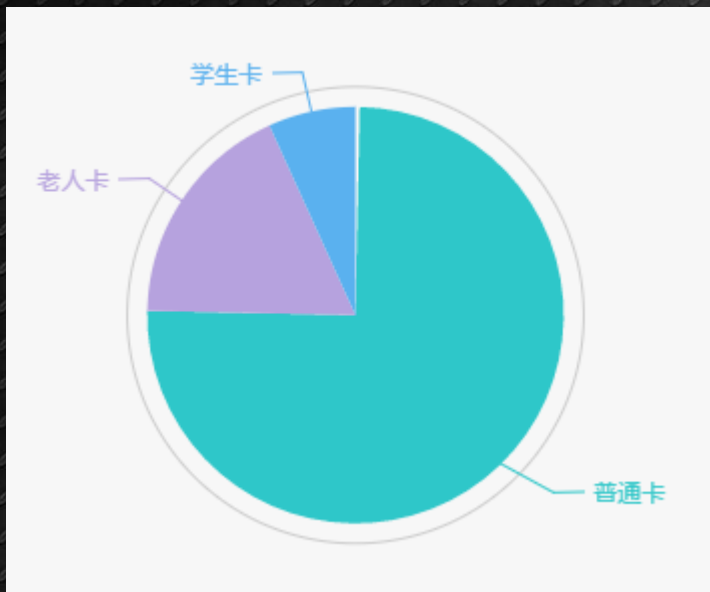
问题分析

数据分析

算法框架

1、乘车类型统计

Part2 --- 乘车总记录数：11596035
乘车总人数：2505883
按时间段统计：6706



分类型乘车比例

卡类型	记录数量	百分比
普通卡	8735614	75.33%
老年卡	2073941	17.88%
学生卡	786480	6.78%

1、其他四种类型中最大比例：0.6%

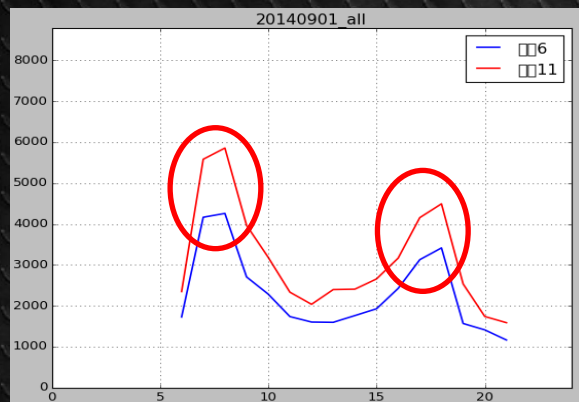
2、将其他四种类型并入普通卡中

问题分析

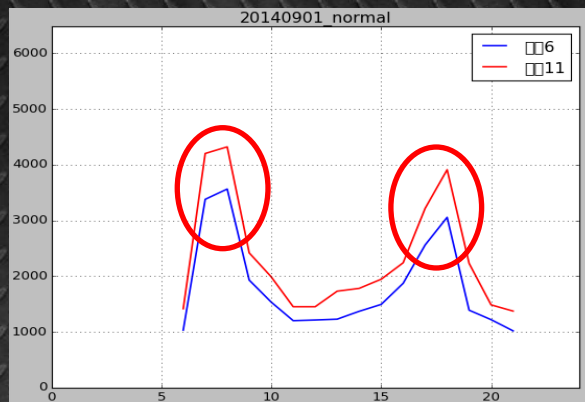
数据分析

算法框架

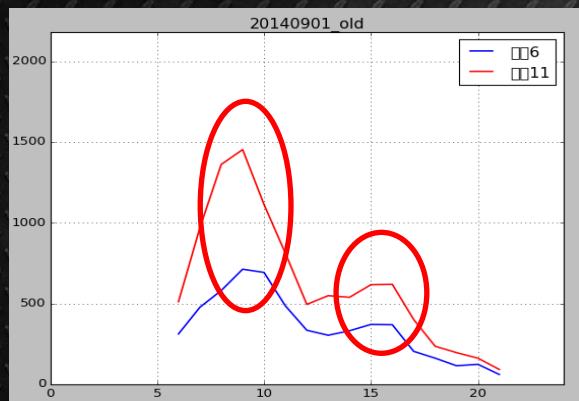
2、工作日波动



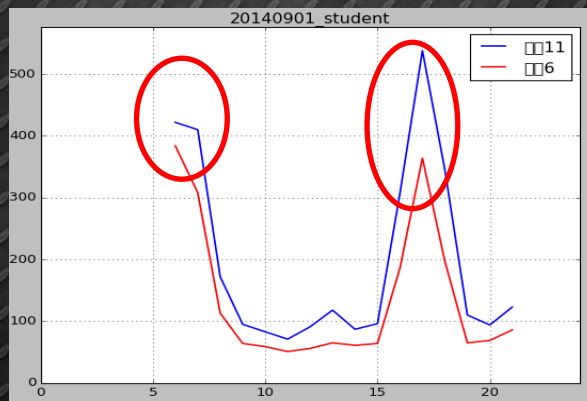
总体--工作日波动



普通卡--工作日波动



老年卡--工作日波动

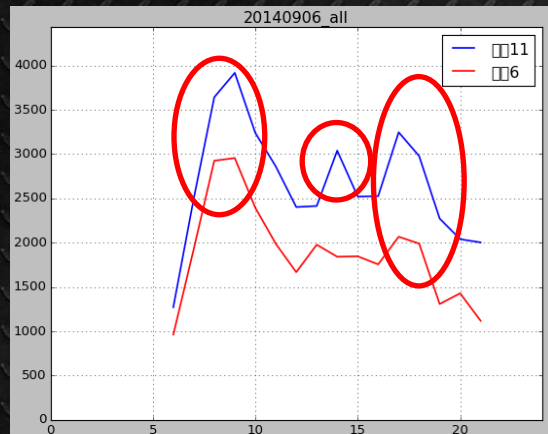


学生卡--工作日波动

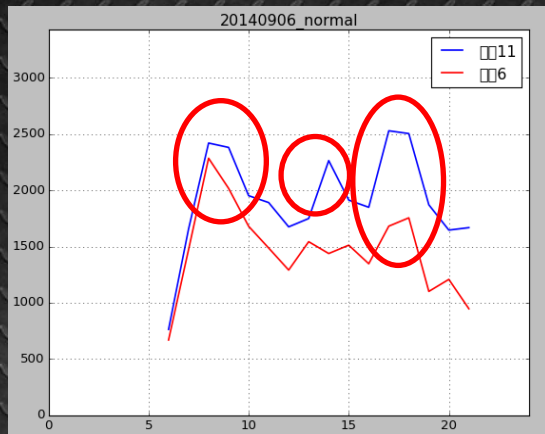
工作日波动情况

- 1、三种人群的波动情况相似
- 2、普通卡对总体影响巨大
- 3、所有工作日波动相似

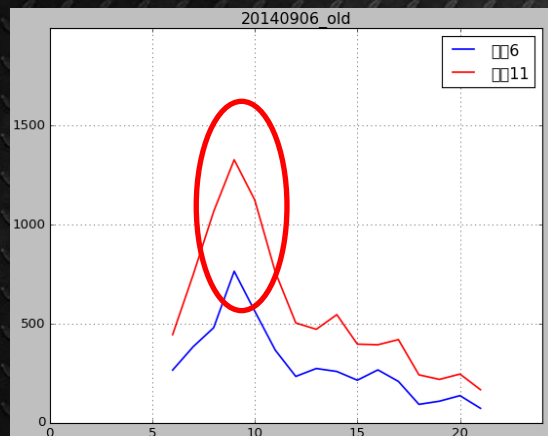
3、周末波动



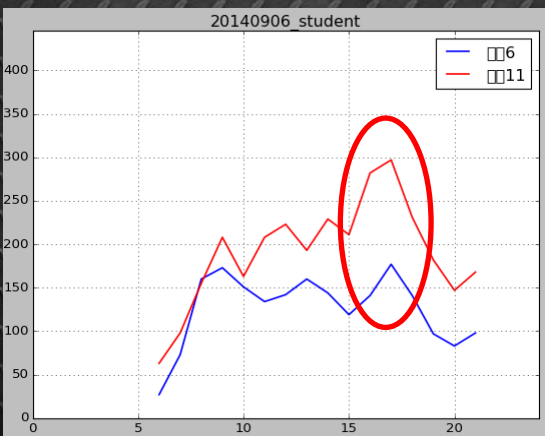
总体一周末波动



普通卡一周末波动



老年卡一周末波动

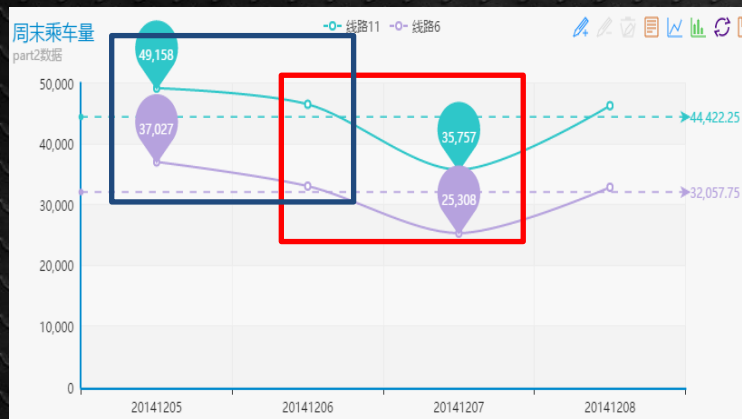


学生卡一周末波动

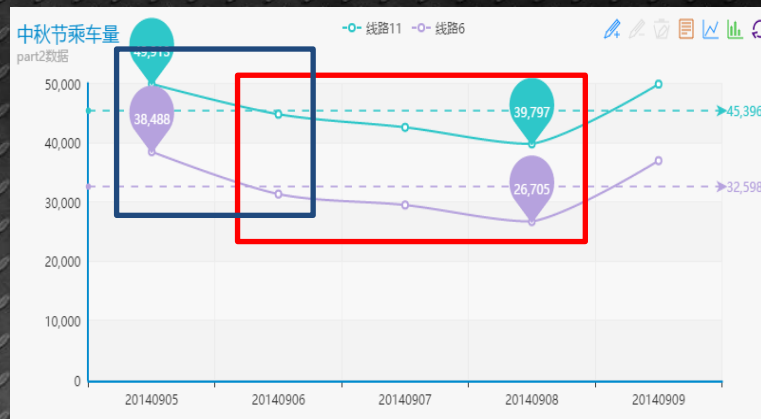
周末波动情况

- 1、普通卡和总体极为相似
- 2、普通卡对总体影响巨大
- 3、老年卡和学生卡只有一个波峰，且出现时间不同
- 4、14时出现了小波峰
- 5、所有周末波动相似

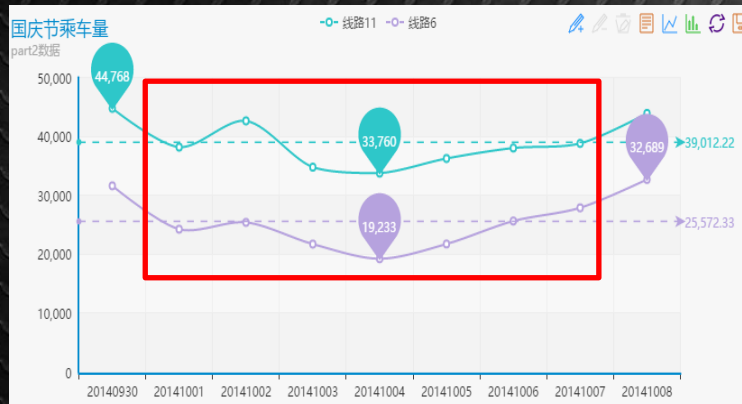
4、节假日波动



周末——乘车记录统计



中秋节——乘车记录统计

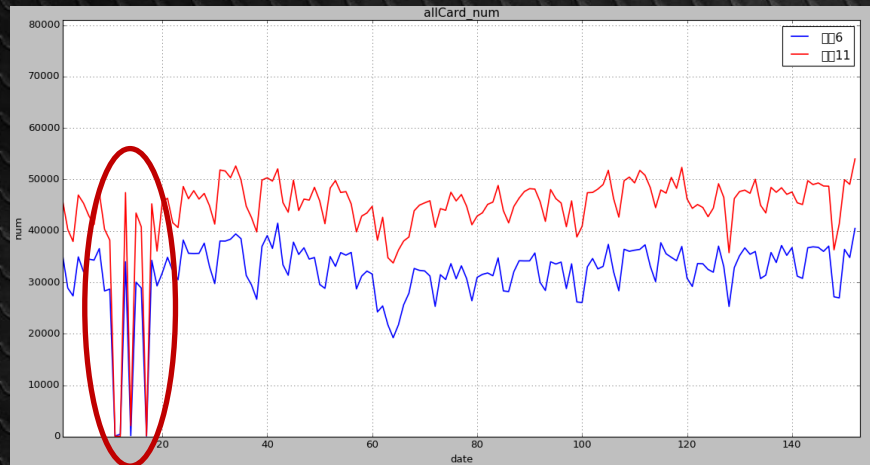


国庆节——乘车记录统计

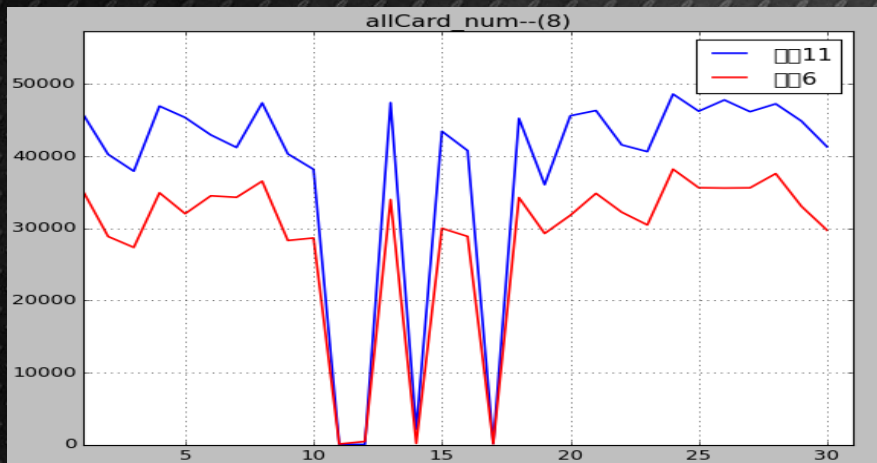
节假日波动情况（总体）

- 1、中秋和周末递减趋势相似
- 2、国庆变化趋势较乱，但有下降趋势
- 3、节假日相比周末下降明显

5、总体波动




8-12月乘车记录统计



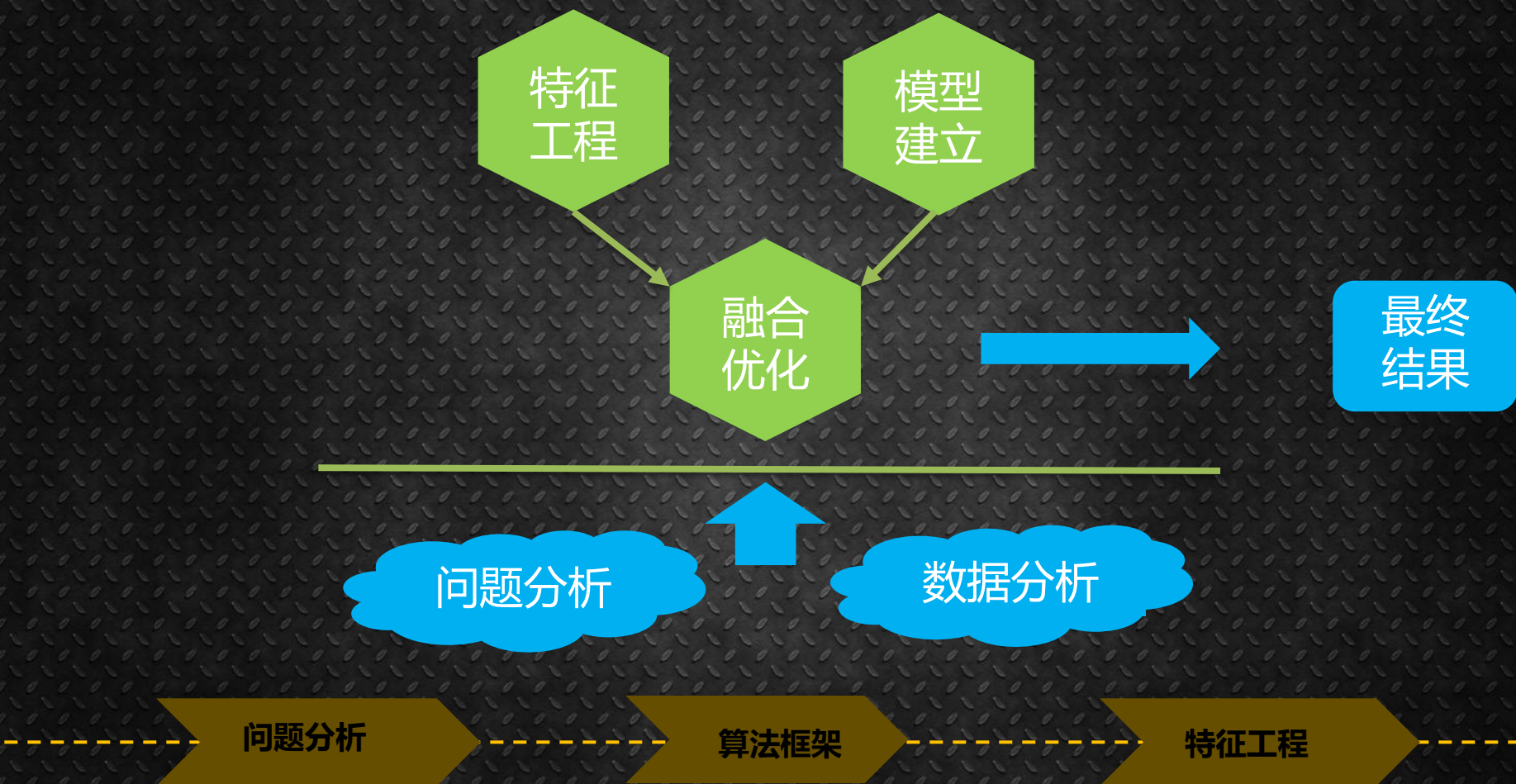
8月乘车记录统计

8-12月波动情况

- 1、在8月份出现异常情况
- 2、8月学生人群总量较其他月份少（1000左右）
- 3、9-12月的各人群波动情况相似



算法 框架



思路 1

分类预测（200万用户可以实现）：
将用户坐车与否转化为二分类问题，针对所有用户得到结果



出现问题

- 1、每日每时乘车人数不确定，分类效果差
- 2、可能对于当前时间会有新用户，出现冷启动情况

问题分析

算法框架

特征工程

思路 2

回归预测：

使用不同人群、不同路线分别建模，最终得到整体回归模型



出现问题

- 1、部分人群数据量小，无法很有效的建模
- 2、最终得到整体回归模型时，整合效果差（误差被叠加）
- 3、不同路线、不同人群建模时可能丢弃了隐含的特征

问题分析

算法框架

特征工程

思路 3

回归预测：
将所有人群、路线整体建模，放弃单独建模



这样做的好处

- 1、建模时，能够有效的保留隐含的一些特征
- 2、数据量大，建模效果好
- 3、单独建模再整合会叠加单独模型中原有的误差，而整体建模不会

问题分析

算法框架

特征工程

算法框架 --- 评测方法



离线评测

1、模型拟合度得分

训练数据在生成模型上的拟合效果，同时模拟线上的预测偏差进行打分

2、交叉评测得分

对训练集合进行交叉验证，得到交叉评分（评测指标多样化）



在线评测

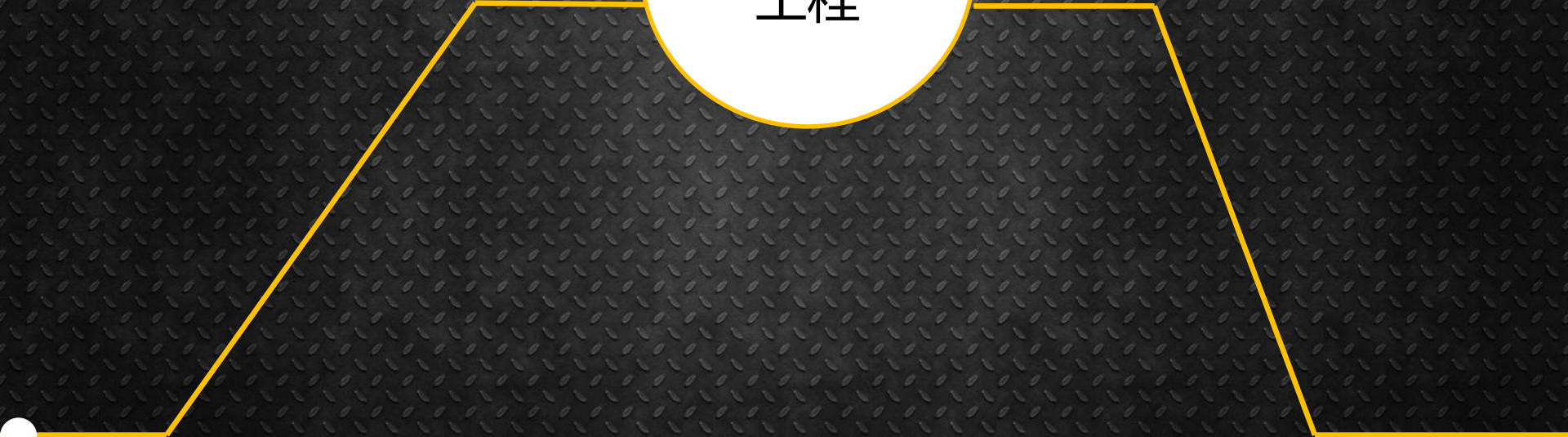
1、平台评分

计算每个预测结果的相对误差，
通过得分函数计算最终的得分

问题分析

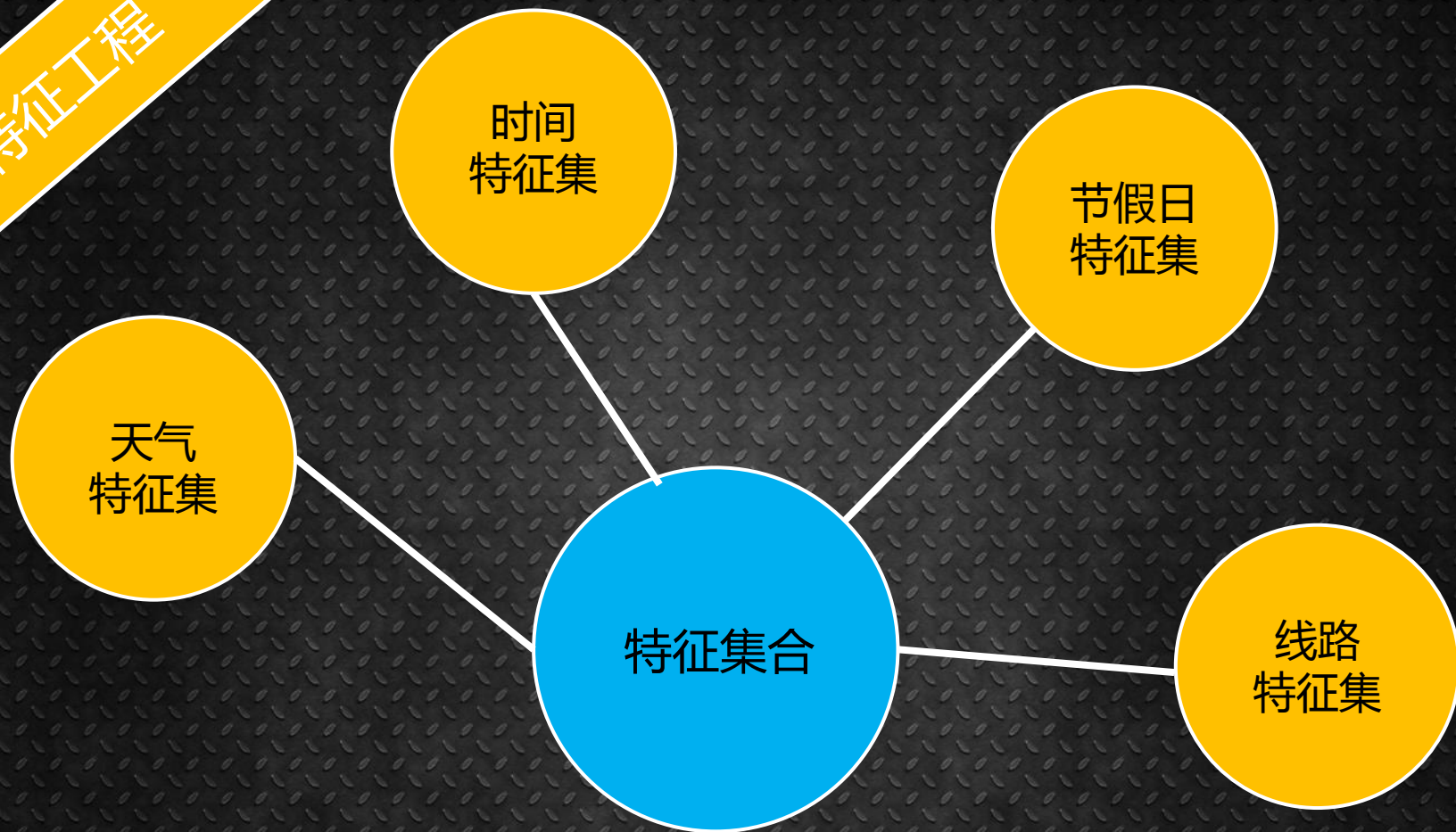
算法框架

特征工程



特征工程

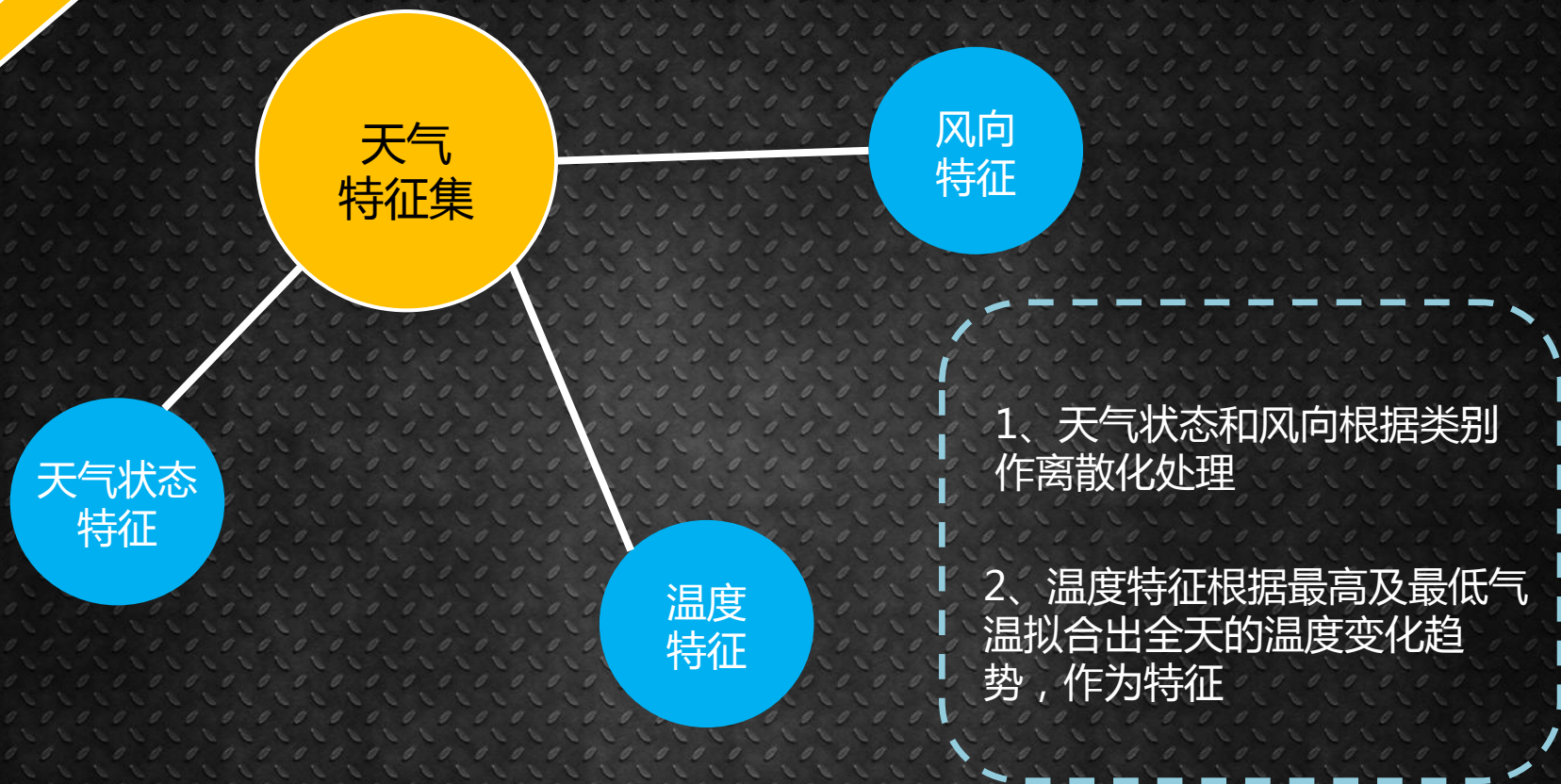
特征工程

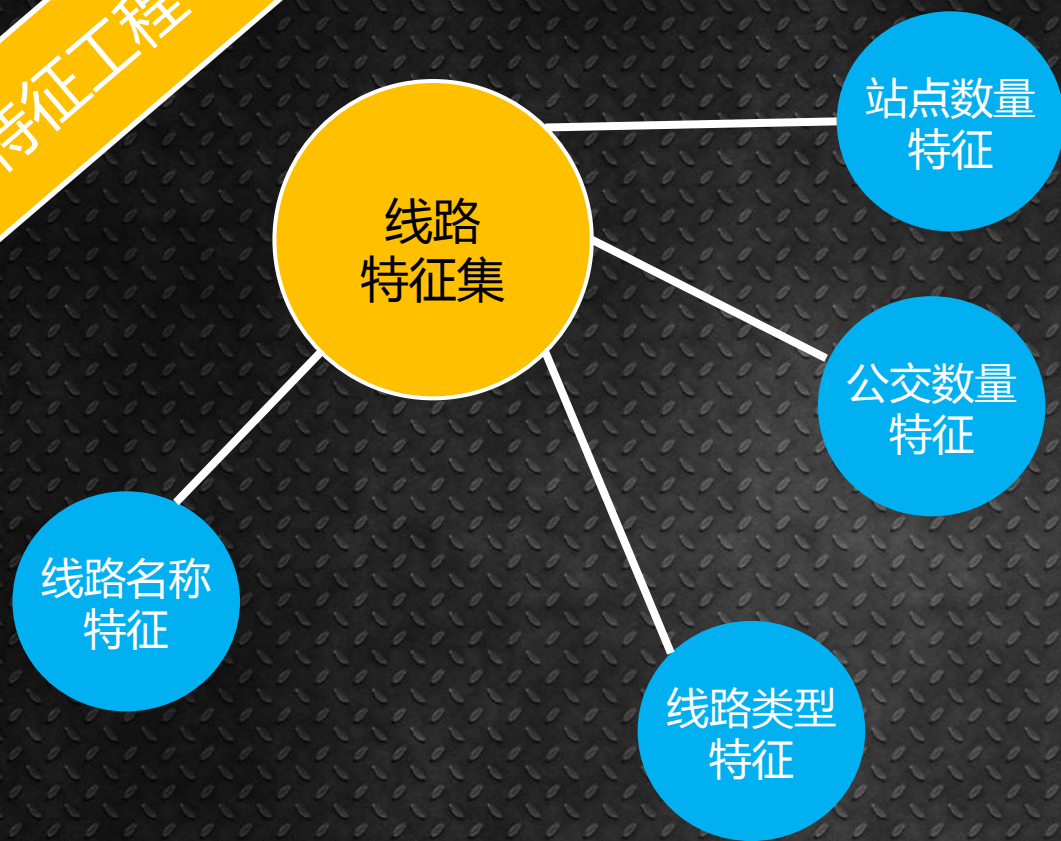


算法框架

特征工程

实证研究





1、线路名称和线路类型进行离散化处理

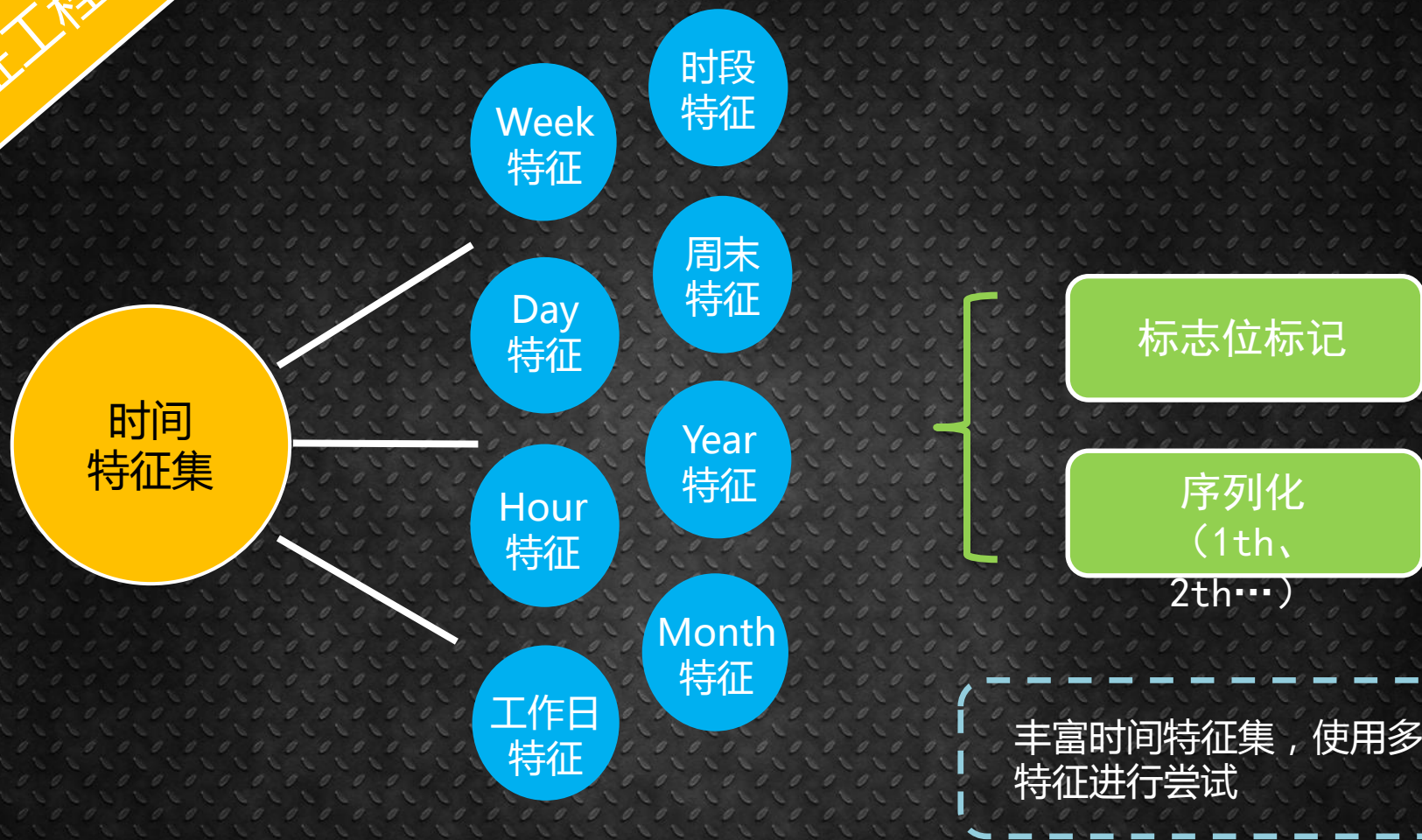
2、站点数量从基本数据直接得到即可

3、不同时段公交数量不同，可以计算后作为特征

算法框架

特征工程

实证研究

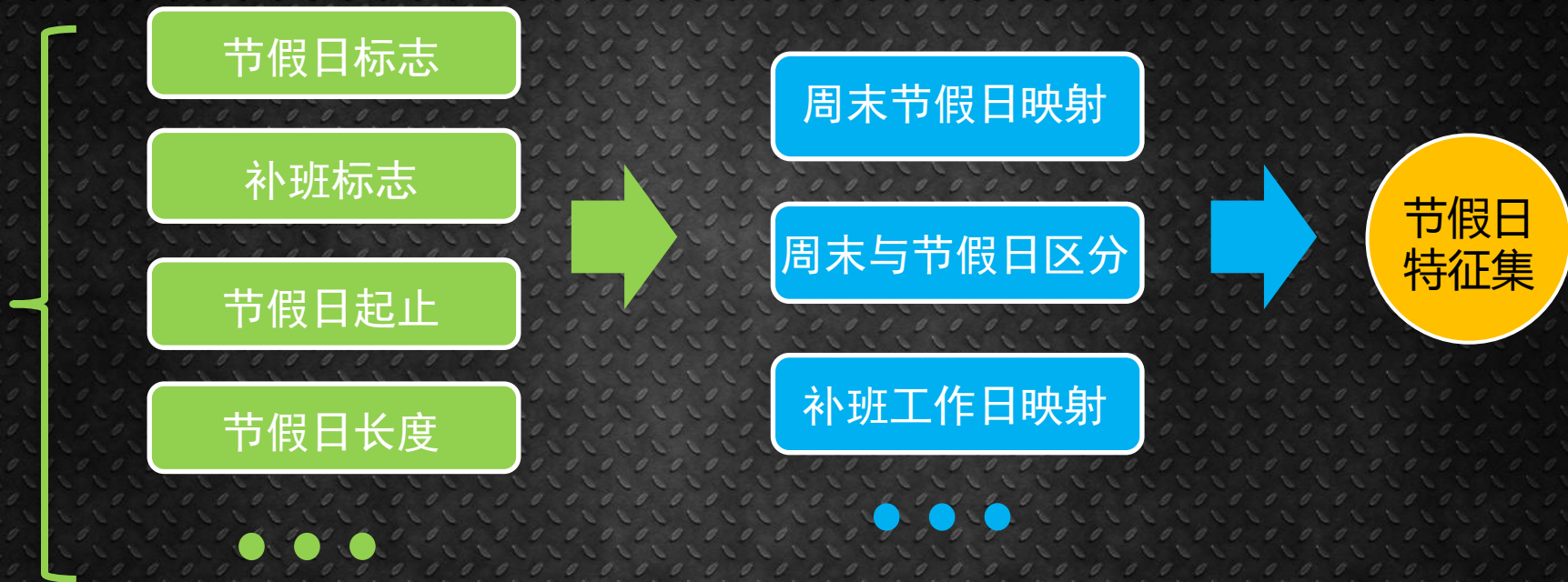


算法框架

特征工程

实证研究

因为1.1—1.7日中有元旦假日的存在，所以引入节假日特征来利用已有的节假日对元旦进行预测。



算法框架

特征工程

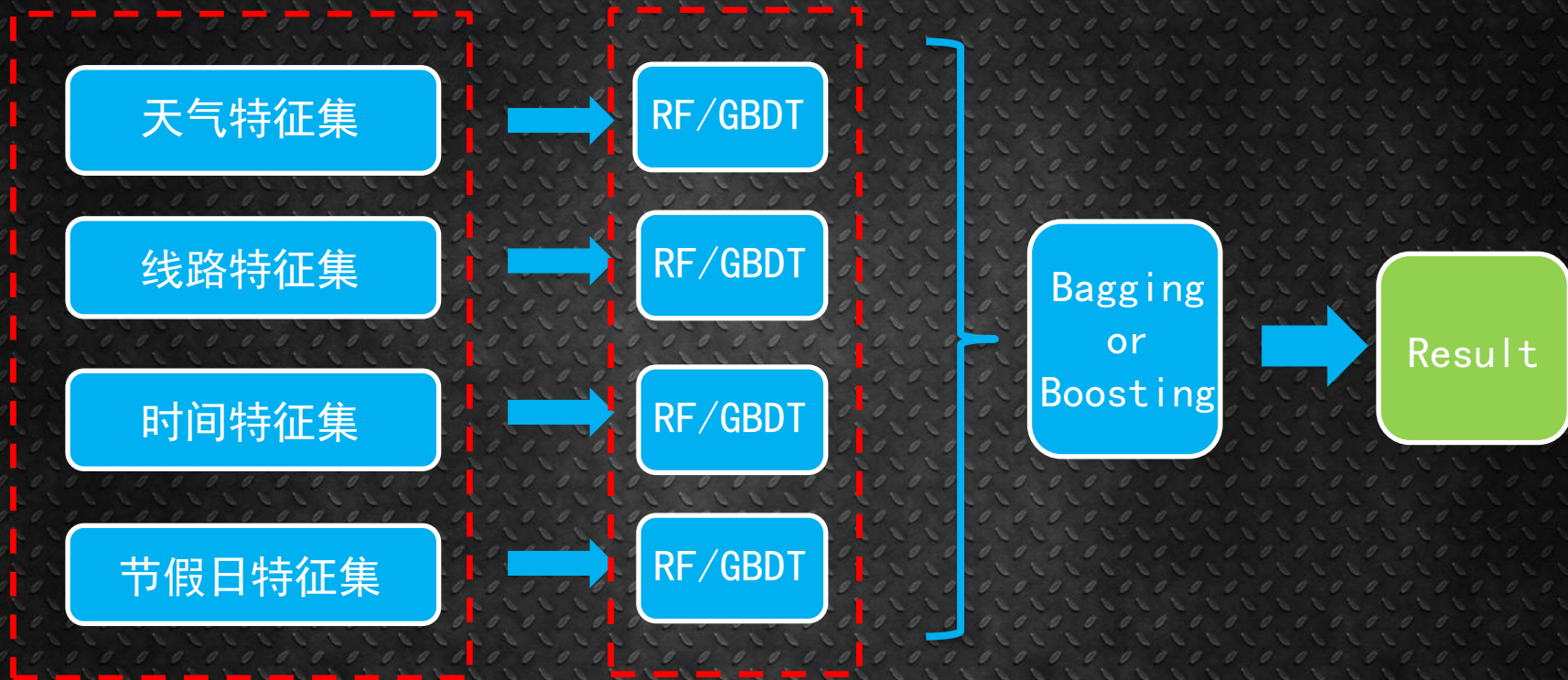
模型建立



模型
建立

模型建立及融合

融合考虑不同时间窗口；不同特征集合；不同训练方式等



特征工程

模型建立

总结

模型评价

首先尝试LR中不同model (ridge...), 效果不理想。
可能原因是数据并不符合线性模型, 无法准确拟合

模型类型	Ensemble方式	模型拟合度	交叉得分	线上得分
LR(Lasso)	Boosting	0.67	0.46	--
SVM	Bagging	0.84	0.72	64.45%
Neural Network	--	0.97	0.96	68.03%
GBDT	Boosting	1	0.95	79.03%
RF	Bagging	0.98	0.91	80.66%

特征工程

模型建立

总结

模型评价

尝试使用SVM模型，进行非线性的拟合，得分有上升，但还是无法得到好的线上得分

模型类型	Ensemble方式	模型拟合度	交叉得分	线上得分
LR(Lasso)	Boosting	0.67	0.46	--
SVM	Bagging	0.84	0.72	64.45%
Neural Network	--	0.97	0.96	68.03%
GBDT	Boosting	1	0.95	79.03%
RF	Bagging	0.98	0.91	80.66%

特征工程

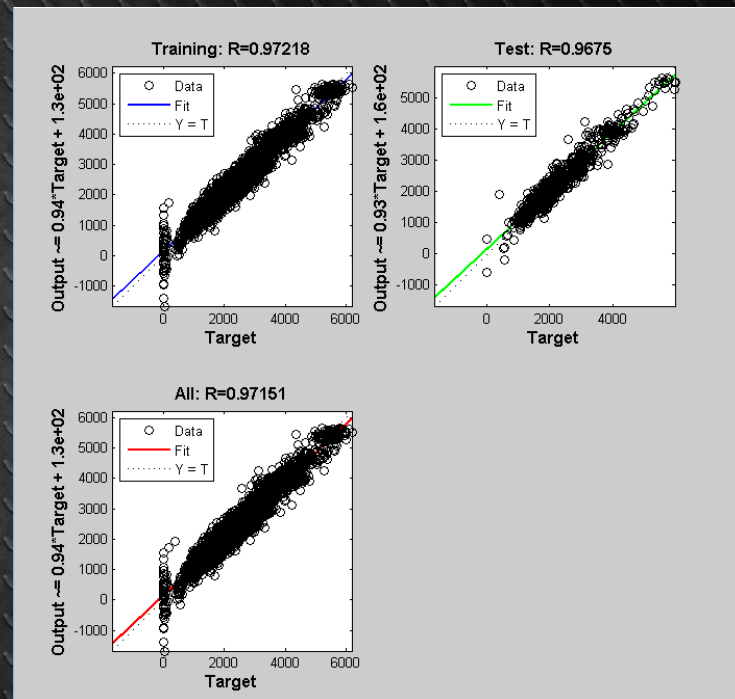
模型建立

总结

模型评价

使用神经网络进行建模，线下效果很好，但会出现不稳定的现象，线上无法得到好的结果（过拟合）

模型类型	Ensemble方式	模型拟合度	交叉得分	线上得分
LR(Lasso)	Boosting	0.67	0.46	--
SVM	Bagging	0.84	0.72	64.45%
Neural Network	--	0.97	0.96	68.03%
GBDT	Boosting	1	0.95	79.03%
RF	Bagging	0.98	0.91	80.66%



特征工程

模型建立

总结

模型评价

使用王牌模型GBDT、RF进行建模，效果很好。在最终的线上评测时，RF分数较高（GBDT速度慢，RF对离散型数据支持较好）

模型类型	Ensemble方式	模型拟合度	交叉得分	线上得分
LR(Lasso)	Boosting	0.67	0.46	--
SVM	Bagging	0.84	0.72	64.45%
Neural Network	--	0.97	0.96	68.03%
GBDT	Boosting	1	0.95	79.03%
RF	Bagging	0.98	0.91	80.66%

特征工程

模型建立

总结

模型评价

*经多次测试，最终选择了RF（Bagging）的方式

模型类型	Ensemble方式	模型拟合度	交叉得分	线上得分
LR(Lasso)	Boosting	0.67	0.46	--
SVM	Bagging	0.84	0.72	64.45%
Neural Network	--	0.97	0.96	68.03%
GBDT	Boosting	1	0.95	79.03%
RF	Bagging	0.98	0.91	80.66%

特征工程

模型建立

总结



总结

总结

- 1、对特征工程中的特征进行测试，过滤得到最优特征，带入回归模型
- 2、数据有取舍的使用（异常值、国庆数据、8月份数据）
- 3、尽可能用特征去拟合曲线的趋势（如假日起始特征）
- 4、深入理解业务、分析数据，特征工程和模型选择时注意细节（参数）

模型建立

总结

End



致谢：

- 1、感谢阿里的天池竞赛
- 2、感谢广东省政府和岭南通公司的精心组织
- 3、感谢所有工作人员细致认真的工作
- 4、感谢所有参赛者，让我收获知识

谢谢

Andy